INFORMATION DISSEMINATION POLICY STATEMENT

EFFECTIVE DATE: July 6, 2005

Subject: Harvesting Federal Digital Publications for GPO's Information Dissemination (ID) Programs

This policy statement governs manual and automated harvesting of publications from Federal agency Web sites for GPO's Federal Depository Library and National Bibliography Programs, information dissemination programs administrated by the Managing Director, Information Dissemination (Superintendent of Documents).

Background

The Government Printing Office (GPO) has been the Government's agent for providing public access to Government information for over a century. The mandates of Chapters 17, 19, and 41 of Title 44, United States Code establish GPO's responsibility for providing permanent public access and comprehensive indexing to tangible and digital Government publications.

GPO defines harvested content as digital content within the scope of dissemination programs that is gathered from Federal agency Web sites. Harvesting technologies are used by GPO to discover and capture publications that have not been cataloged by GPO but fall within the scope of the Federal Depository Library Program (FDLP) and the National Bibliography.

Guided by 44 U.S.C. §§1901-1903, GPO's FDLP and the National Bibliography are focused upon the final, published versions of agency publications, in this case, those that are published on the Web.

Policy

GPO will acquire online publications for inclusion in the National Bibliography and the Federal Depository Library Program through manual and automated harvesting. GPO will use automated harvesting programs only with the publishing agency's advice and prior consent. Permission to manually harvest publications from agency publicly accessible Web sites will not be sought.

I. Advice and Consent from Publishing Agencies

A. The Superintendent of Documents will contact, in writing (e-mail), the agency's Chief Information Officer. The letter will:

- Explain why GPO is harvesting agency Web sites;
- Describe how the harvested files will be used;
- Seek the agency's advice and consent for GPO to employ an automated harvesting application to their Web site;

- Seek the agency's advice for frequency and timing of harvests;
- Convey to the agency that manual harvesting of individual publications that are on publicly available Web sites will be conducted if permission to use automated harvesting is denied;
- Ask the agency to share with GPO a copy of any harvesting policies they have in place; and
- Communicate that if, after 30 days, the target agency has not responded to the letter, the non-response will be considered consent to harvest.

B. GPO will manually harvest individual publications that reside on publicly accessible Web sites.

II. Harvesting

A. GPO will follow security and industry best practices to ensure minimal impact on Federal agency (target) Web servers as automated and manual harvesting is employed.

B. GPO will honor harvesting protocols established by the publishing agencies
   a. Notices on Web sites
   b. Robot exclusions

C. Automated harvesting program will indicate that GPO is harvesting and contact information for the Acquisitions and Development Director will be included.

D. Agency Web sites will be reharvested on a schedule determined by the Director, Acquisitions & Development and that is consistent with the advice of the publishing agency.

E. Any automated tools or services used for harvesting by GPO must be approved by GPO's Office of the Chief Information Officer (OCIO).

III. Harvested Files

A. Among the metadata elements of harvested files will be language that indicates it is an authorized, captured, and archived file of the original.

B. An automated or manually harvested publication may be determined to be out of scope at the time it is cataloged. If indicators such as "For Official Use Only" or "Restricted" appear on the publication, the agency will be contacted by Office of Bibliographic Services staff for verification of status.

C. Out of scope files acquired via automated or manual harvesting will be deleted from GPO servers. The files will never have been accessible by the public through GPO.

D. All harvested publications are subject to recall under ID 72, Withdrawal of Federal Information Products from GPO's Information Dissemination (ID) Programs.

Limitations

This policy pertains to all U.S. Government information products and services subject to the jurisdiction of the Superintendent of Documents. However, the following limitations apply:

- GPO harvests from official Web sites, the originating agency or its business partner(s). An example of an agency business partner is the University of Albany, School of Criminal Justice, which hosts the Bureau of Justice Statistics' annual Sourcebook of Criminal Justice Statistics (http://www.albany.edu/sourcebook/).
- GPO also will manually harvest in-scope publications from unofficial sources, such as institutions creating digital access copies or non-Governmental Internet archives, when they are no longer available from official Government Web sites and GPO does not have an archived copy. GPO provides digital content with varying levels of authentication dependant upon provenance, chain of custody, and level of quality assurance in or type of output from a legacy digitization process. Publications harvested from unofficial sources are considered low-confidence access copies.

Application

This policy applies to all appropriate elements of Information Dissemination. The Managing Director, Information Dissemination, Superintendent of Documents must authorize any exceptions to this policy. Exceptions will be documented in writing to the Acquisitions & Development staff.

Referenced Policies

ID 72: Withdrawal of Federal Information Products from GPO's Information Dissemination (ID) Programs

Approved _____

Managing Director, Information Dissemination
(Superintendent of Documents)